DOCUMENT RESUME

ED 046 195 EC 031 391

AUTHOR Bryant, N. Dale
TITLE Correcting Correlations for Restriction in Range Due to Selection on an Unmeasured Variable. Interim Report.
INSTITUTION Columbia Univ., New York, N.Y. Teachers College.
SPONS AGENCY Bureau of Education for the Handicapped (DHEW/OE), Washington, D.C.
BUREAU NO BR-42-2001
PUB DATE Jun 70
GRANT OEG-2-7-070701-4249
NOTE 12p.

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Comparative Analysis, *Correlation, *Research Methodology, *Sampling, *Statistical Analysis, Validity
IDENTIFIERS Pearson Product-Moment Correlation

ABSTRACT

The paper points out that, in working with special groups, correlations are often distorted because the variability of the measures being correlated are restricted in the groups. Presented is a formula whereby a Pearson product-moment correlation can be corrected for restrictions in range in situations where the basis of selection is unmeasured, but where the extent of restriction for each of the two measures being correlated is known, and where the variables are assumed to be normally distributed in the population. Three examples of the use of the formula are given: in a case where a comparison is to be made between a value derived from an unrestricted sample and one derived from a restricted sample; a case when a correlation is obtained on a special restricted sample and must be generalized to the population; and in estimating the validity of a test, where the criterion and the test scores are available on the same individuals only in a restricted sample where the basis of the selection is not clear or not measured. (Author/KW)

INTERIM REPORT

Project No. 422001
Grant No. OEG-2-7-070701-4249

CORRECTING CORRELATIONS FOR RESTRICTION IN RANGE DUE TO SELECTION ON AN UNMEASURED VARIABLE

N. Dale Bryant
Research and Demonstration Center
for the Education of Handicapped Children
Teachers College, Columbia University
New York, New York

June 1970

Department of Health, Education, and Welfare

U.S. Office of Education
Bureau of Education for the Handicapped

INTERIM REPORT

Project No. 422001
Grant No. OEG-2-7-070701-4249

CORRECTING CORRELATIONS FOR RESTRICTION IN RANGE DUE TO SELECTION ON AN UNMEASURED VARIABLE

N. Dale Bryant
Teachers College, Columbia University
New York, New York
June 1970

The research reported herein was performed pursuant to a grant with the Bureau of Education for the Handicapped, U.S. Office of Education, Department of Health, Education, and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official positions of the Bureau of Education for the Handicapped.

Department of Health, Education, and Welfare

U.S. Office of Education
Bureau of Education for the Handicapped

1970 - 1

ABSTRACT

In working with special groups, correlations are often distorted because the variability of the measures being correlated are restricted in the groups. The formula presented in this paper can be used to correct product-moment correlations for this distortion even when the basis of the restriction is unknown.

CORRECTING CORRELATIONS FOR RESTRICTIONS IN RANGE DUE TO SELECTION ON AN UNMEASURED VARIABLE*

H. Dale Bryant
Teachers College, Columbia University

Sunanda Gokhale
Albany, New York

The size of a correlation coefficient is dependent in part upon the variability of the measured values in the correlation sample. Any time that a sample is restricted in range on either or both of the measures, the correlations between those two measures will tend to be lowered as compared to the same correlation based upon a representative sample of the population. If prediction within the restricted sample is the purpose of the correlation, then the obtained value is the meaningful and correct one. However, if, for some reason, it is not possible to correlate the variables using an unrestricted sample, we can infer the relationship between the two measures irrespective of the restriction if we correct the correlation for the effect of the restriction in range. For example, if, in a sample of bright students, reading achievement and academic grades show only a .2 correlation, we cannot infer that this is the general relationship between reading and school grades. Since a high IQ group will tend to make high grades and will also tend to be high on reading ability, there is likely to be severe restriction in range on both variables. For prediction within the high IQ group, the .2 correlation is appropriate, but to infer beyond the sample, a correction for restrictions in range is necessary. Guilford (1965, pp. 341-345) gives three formulae, attributed to Karl Pearson, to correct a Pearson product-moment correlation coefficient for restriction in range when restriction results from selection on one of the two variables being correlated or on some measured

*The work presented or reported herein was performed pursuant to a grant from the U.S. Office of Education, Department of Health, Education and Welfare.

third variable. The assumption must be made that the variables are normally distributed in the population.

PROBLEM

In many clinical and other settings, the sample is obviously restricted in range on different variables, but the basis for the restrictions (i.e., the selection variables) is complex, unknown, or unmeasurable. Examples of such sampling might be children coming to a particular clinic, cases receiving a particular diagnosis, or individuals exhibiting a particular behavior. In all these cases, the samples may show restrictions in range on variables being correlated, but the basis of the restrictions cannot be reduced to a measurable variable. In these instances, the formulae presented by Guilford cannot be used. It is possible, however, to correct for restrictions in range, even though the selection variable is unknown or unmeasured, by using information about the extent of the restriction on each of the two variables being correlated.

This paper presents a formula whereby a Pearson product-moment correlation can be corrected for restrictions in range for these special but very frequent situations where the basis of selection is unmeasured but where the extent of restriction for each of the two measures being correlated is known and where the variables are assumed to be normally distributed in the population.

FORMULA FOR USE WHEN RESTRICTIONS RESULT FROM COMPLEX OR UNMEASURED VARIABLES

Starting with Guilford's formula for correcting r_{12} for restriction in range, we can rewrite his Formula II so that it corrects a correlation r_{31}, where restriction is produced by selection on the basis of variable 3 and there is knowledge of the standard deviations for variable 1 in both the restricted and unrestricted samples. Similarly, we can rewrite his Formula I so that it corrects a correlation r_{31}, where restriction is produced by selection on the basis of variable 3 and there is knowledge of the standard deviations for variable 3 in both the restricted and unrestricted groups. By equating these two formulae and squaring and simplifying them, we can obtain an equivalent value for the ratio of unrestricted to restricted variances on variable 3, expressed in terms of the ratio of unrestricted to restricted variances on variable 1 and the correlation r_{31}. The same procedure can be followed by rewriting Formulae I and II to correct r_{32} so as to obtain an equivalent value for the ratio of unrestricted to restricted variances on variable 3, expressed in terms of the ratio of unrestricted to restricted variances on variable 2 and the correlation r_{32}. Thus, the information about restriction on variable 3 is expressed in terms of information about the variables 1 and 2 and the correlations r_{31} and r_{32}.

These equivalent ratio values described above can be substituted into Guilford's Formula III (for R_{12}), where restriction is produced by selection on the basis of variable 3 and there is knowledge of the standard deviations for variable 3 in both the restricted and unrestricted groups and where r_{13} and r_{23} are known. However, since there are two estimates of the ratio of unrestricted to restricted variances on variable 3, we must express the value as the square root of the product of the two estimates (viz., $a = \sqrt{a \times a}$).

The resulting formula for the corrected correlation (R_{12}) is given below:

$$R_{12} = r_{12}\sqrt{\frac{s_1^2}{\sigma_1^2} \times \frac{s_2^2}{\sigma_2^2}} + \sqrt{\left(1 - \frac{s_1^2}{\sigma_1^2}\right)\left(1 - \frac{s_2^2}{\sigma_2^2}\right)}$$

This formula does not require all of the information necessary for Guilford's Formulae I, II, and III, but it can be used to obtain a product-moment correlation coefficient that is corrected for restrictions in range (R_{12}) knowing only the uncorrected correlation (r_{12}), the standard deviations of the two variables in the restricted samples (s_1 and s_2), and the standard deviations of the two variables in the unrestricted sample (σ_1 and σ_2).[1]

EXAMPLES OF USE OF THE FORMULA

In a clinical sample of children, it was noted that a particular measure (the Coding subtest on the Wechsler Intelligence Scale for Children) was consistently lower than the average of the other intelligence subtests. The sample consisted of children of average or above averabe IQ who were brought by their parents to a clinic because school remedial procedures were not correcting the children's severe reading retardation. To study the nature of this lowered performance, the variable, Coding, was correlated with other reference variables such as the Perceptual Speed Test of the Primary Mental Abilities Test Battery. The correlation of a reference variable and the Coding subtest needs to be compared to equivalent values in a sample representative of the population as given in other research studies. In order to make the correlation based upon the clinical sample comparable to the correlation based upon the sample representative of the population, it is necessary to correct for restrictions in range, since both Coding and the reference variable, Perceptual Speed, show consistently lower scores than are normally found in a presumably representative sample from the population. The specific factors responsible for the restriction in range cannot be measured, since coming to a clinic involves much more than poor reading. In both Coding and Perceptual Speed, we can assume normality of distribution within the population.

The values obtained for the clinic sample are as follows: $r_{12} = .40$, where 1 and 2 represent Coding and Perceptual Speed respectively; $s_1^2 = 2.59$ and $s_2^2 = 186$, where s^2 is the variance for the clinic sample. Equivalent values for normative samples of appropriate age as given in the manuals for the respective tests are $\sigma_1^2 = 9$ and $\sigma_2^2 = 289$, where σ^2 is the variance based upon the normative samples. Substituting in the final formula given above:

$$R_{12} = .40\sqrt{\frac{2.59}{9} \times \frac{186}{289}} + \sqrt{\left(1 - \frac{2.59}{9}\right) \times \left(1 - \frac{186}{289}\right)} = .68$$

A study, based upon a "normal" sample of eighth grade children (which is roughly comparable to the grade placement of the clinic sample) and having variances similar to the population values, reported that $r_{12} = .37$.

By using the correction for restrictions in range, it is possible to compare the .68 in the clinic sample with the .37 in the normal sample. It suggests that there is a higher degree of relationship between these two measures in the clinic sample (and confirms certain conclusions drawn from clinical observation). While it is beyond the scope of this paper to comment upon the interpretation of this finding, it is apparent that interpretations could be made that could not have been made if there had been no correction for restrictions in range.

The illustration above is of a case where a comparison is to be made between a value derived from an unrestricted sample and one derived from a restricted sample. The values have to be expressed in comparable terms, so the correction for restrictions is necessary.

Another example of a case where the correction for restrictions in range is necessary is when a correlation is obtained on a special, restricted sample and must be generalized to the population. An example of this might be a study of the relationship between the amount of a particular chemical in the blood and the frequency of hallucinatory-type activity. Since this is hard to study in a nonclinical population, we might study it in a sample of individuals diagnosed as schizophrenic. If schizophrenics seldom have a low concentration of the chemical in their blood and if they tend to show more frequent hallucinatory-type activity than would be true for the total population, then both of these variables are restricted in range. A correlation between the two variables in the schizophrenic sample can be used to infer what the relationship would be in the total population <u>if it is assumed</u> that the same relationship holds true for lower levels of the chemical and less frequent hallucinatory-type activity and that the clinical sar ･ nerely represents one end of a distribution on these two variables, which ar normally distributed in the population. While these assumptions might not be justified, it is evident that, if they are made, the correlation based upon the s izophrenic sample would have to be corrected for restrictions in range i order to infer the relationship in the population. The basis of the selection o the sample is complex, and, unless a measure of the selection variable ca be obtained, it would be necessary to use a formula such as the one presented i nis paper.

Anoth example of the application of the formula would be its use in estimating t e validity of a test where the criterion and test scores are available on the sa individuals only in a restricted sample where the basis of the selection is not clear or not measured. If the variance of the test is known for some sample tha is representative of the population and the variance of the criterion is known r some other sample representative of the population, the formula can provide orrection to estimate the validity of the test in an unrestricted sample.

SUMMARY

There are many times that Pearson product-moment correlations are based on clinical samples or other special groups where there are restrictions in range on the variables being correlated and where the basis of the selection that causes the restrictions is unknown or unmeasured. It is often necessary either to compare the correlation with values derived from a sample representative of the population or to infer from the special sample the nature of the relationship that exists between the two variables within the total population. In such cases, if the assumption can be made that the variables are normally distributed in the population, the formula presented in this paper is applicable in correcting the correlation coefficient for restrictions in range.

REFERENCE

Guilford, J. P. _Fundamental statistics in psychology and education._ New York: McGraw-Hill, 1965.

FOOTNOTE

[1]In kindly checking this derivation, Dr. Rosedith Sitgreaves, Principal Advisor, Educational Research and Statistical Methods Area, Psychology Department, Teachers College, Columbia University, pointed out that the formula could be obtained somewhat more directly without recourse to the Guilford formulae. The senior author will be happy to send upon request both the original and Dr. Sitgreaves' derivations to anyone requesting them.